International Journal of Cyber-Physical Systems
Volume 2 • Issue 1 • January-June 2020

Exploring the Potential of an Extensible Domain-Specific Web Corpus for "Layfication":
The Case of Cross-Lingual Classification

Marina Santini, RISE Research Institutes of Sweden, Sweden
Min-Chun Shih, Linköping University, Sweden

ABSTRACT

This article presents experiments based on the extensible domain-specific web corpus for "layfication". For these experiments, both the existing layfication corpus (in Swedish and in English) and a new addition in English (the NHS-PubMed subcorpus) are used. With this extended corpus, methods to classify lay-specialized medical sublanguages cross-linguistically using small data and noisy web documents are investigated. Sublanguage is a language variety used in specific domains. Here, the authors focus on two medical sublanguages, namely the "patientspeak" (lay) and the medical jargon (specialized). Cross-lingual sublanguage classification is still largely underexplored although it can be crucial in downstream applications for digital health and cyber-physical systems. Classification models are built using small and noisy training sets in Swedish and evaluated on English test sets. The performance of Naive Bayes classifiers—built with stopwords and with Bag-of-Words—is compared with convolutional neural network classifiers leveraging on MUSE multi-lingual word embeddings. Results are promising and nuanced. These results are proposed as a first baseline for cross-lingual sublanguage classification.

KEYWORDS

CNN, Convolutional Neural Network, Medical Domain, Naïve Bayes, Noise, Noisy Data, Small Data, Specialized Corpus, Sublanguage, Web Corpus

DOI: 10.4018/IJCPS.2020010102

INTRODUCTION

Cyber-Physical Systems (CPSs) belong to an emergent paradigm that combines most advanced technological approaches and computational tools to solve complex tasks. CPSs are domain-independent and have penetrated diversified disciplines, such as healthcare and self-driving vehicles.

In the era of data-driven science, corpus-based Language Technology is an essential component of many CPSs, where linguistic knowledge is indispensable to prevent failures or fatal errors due to misunderstandings or poor understanding during human communication, especially when it comes to the interaction between doctors and patients. Arguably, Language Technology can help because it empowers patients and other non-professional actors to understand medical information. Since modern Language Technology and Artificial Intelligence are data-driven, the need of representative text corpora is paramount. Curated, big and "clean" corpora are expensive and beyond the budget of practical domain-specific medical applications. Rather, real-world data is often small and noisy but nonetheless potentially useful. To address this issue, a textual resource was created in the form of an extensible web corpus for "layfication" (Santini et al. 2019) whose main purpose is to explore and test methods based on small and noisy data. The aim is to use the extensible web corpus for layfication as a sandbox to explore the benefits and the limitations of computational methods that could enhence digital health and Cyber-Physical Systems.

In Santini et al. (2019), a corpus for layfication was been presented and made available to the community for re-use, enhancement and expansion. The corpus was conceived as a flexible and extensible textual resource, where additional documents and additional languages can be appended over time. The corpus was started with web pages collected from the internet and includes documents describing a number of diseases in lay and specialized sublanguages. Although progressively expanded, the corpus is still small, because information about some diseases is not extensive. This paucity of textual data in some areas is a real-world condition. The main purpose of the corpus is to be used for building and training language technology applications for the "layfication" of the specialized medical jargon, where "layfication" refers to the automatic identification of more intuitive linguistic expressions that can help laypeople (e.g., patients, family caregivers, and home care aides) understand medical terms, which often appear opaque. In the medical field, layfication is also denoted by the expression "patientspeak", to indicate that the layfication is needed by many patients in their interaction with healthcare professionals.

Why is layfication needed? The medical domain hinges upon medical terminology elaborated and used by healthcare professionals. Despite it is widely acknowledged that understanding what the doctors say has an important influence on the success of treatments, in many cases medical terminology hinders the comprehension of various groups of people (such as non-native speakers, people with low education, etc.), and has negative effects on health consumers (e.g. patients and caregivers). As a matter of fact, in the medical domain, two broad user groups interact. The first group (the expert)